DOCUMENT RESUME

ED 430 095 CE 078 565

AUTHOR Raymond, Mark R.
TITLE Job Analysis and the Development of Test Specifications for Licensure and Certification Examinations.
PUB DATE 1995-04-00
NOTE 36p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 1995).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Certification; Credentials; Data Analysis; Data Collection; Information Sources; *Job Analysis; *Licensing Examinations (Professions); Specifications; Task Analysis; *Test Construction; Test Items; *Test Validity; *Vocational Education

ABSTRACT

This paper reviews and evaluates methods for conducting job analysis. The paper begins with a discussion of the purpose of licensure and certification and the dimensions of job analysis. The following four dimensions along which job analysis methods vary are discussed: types of job descriptors; sources of information; data collection methods; and level of detail or analysis. The next four sections explore the following issues: selecting the most effective methods for conducting job analysis for credentialing examinations (task inventory, professional practice model, critical incident technique, functional job analysis, combined methods); deciding whether descriptors for characterizing occupations and professions should be task oriented or worker oriented; determining how knowledge, skill, and ability domains should be specified (surveying practitioners about knowledge, skills, and abilities; empirical methods for linking knowledge, skills, and abilities; and subject-matter expert panels for judging knowledge, skills, and abilities); and deciding how job analysis results should be translated into test specifications (the challenge, organizational frameworks, and weights for test specifications). Concluding the paper is an examination of the following considerations in evaluating the effectiveness of job analysis: protecting the public; protecting examinees; protecting credentialing agencies; and weighing the science and politics of validity. The paper contains 80 references. (MN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 095

CE

Job Analysis and the Development of Test Specifications for Licensure and Certification Examinations

Mark R. Raymond

The American Registry of Radiologic Technologists

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the National Council on Measurement in Education

San Francisco, CA

April, 1995

Running head: Job Analysis and Test Specifications

CE 070 565

BEST COPY AVAILABLE

Contents

Introduction 2

Rationale for Job Analysis 2
Dimensions of Job Analysis 3

Research Questions 4

Question #1: *Which methods for conducting job analysis for credentialing exams are most effective?* 5

Task Inventory 5
Professional Practice Model 7
Critical Incident Technique 8
Functional Job Analysis 9
Combined Methods 10

Question #2: *Should descriptors for characterizing occupations and professions be task oriented or worker oriented?* 11

Task–Oriented Descriptors 11
Person–Oriented Descriptors (KSAs) 12

Questions #3: *How should KSA domains be specified?* 13

Surveying Practitioners about KSAs 13
Empirical Methods for Linking Tasks and KSAs 16
SME Panels for Judging KSAs 17

Question #4: *How should the results of a job analysis be translated into test specifications?* 19

The Challenge 19
Organizational Frameworks 20
Establishing Weights for Test Specifications 23
Organizational Frameworks Revisited 25

Evaluating the Effectiveness of Job Analysis 28

Protecting the Public 28
Protecting Examinees 29
Protecting Credentialing Agencies 29
The Science and Politics of Validity 30

References 31

Job Analysis and the Development of Test Specifications for Licensure and Certification Examinations

Introduction

Rationale for Job Analysis

The purpose of licensure and certification is to protect the public by assuring that individuals who work in an occupation or profession are at least minimally competent (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 1985). To be consistent with this purpose, licensure and certification examinations must be job related. The job relatedness of a credentialing[1] exam is typically assured by developing test specifications based on a job analysis. Job analysis is advocated by most experts in credentialing (AERA, et al., 1985; D'Costa, 1986; Kane 1982; Shimberg, 1981; Smith & Hambleton, 1990), and has been supported by the courts and federal guidelines (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978, Kuehn, Stallings, & Holland, 1990; Thompson & Thompson, 1982)

A job analysis for a credentialing exam can be conducted using of a wide variety of data collection and analysis procedures commonly used in the social sciences. One of the most common approaches to job analysis is to administer a survey to determine those job activities considered important for a particular occupation (Gael, 1988). Most surveys consist of task inventories, although some have focused on knowledge, skills, and abilities (KSAs). Numerous other methods of job analysis have been used, as well. For example, the critical incident technique (Flanagan, 1954) has been popular, especially in medicine. There has been no shortage of methods. While many have been detailed and rigorous, others have been vague and incomplete. Shimberg (1990) commented on the substantial diversity in methods for conducting a job analysis and called for standards to provide formal guidance concerning how a job analysis should be conducted in order to produce valid test specifications or test plans. The variation in job analysis is highlighted in a recent article by Nelson (1994), who studied the job analyses and test plans developed by two different state jurisdictions for a the same occupation. Although both states had conducted a job analysis of this particular occupation, the two test plans had only about 50% in common in terms of content (Nelson, 1994, p. 32). One explanation for the differences in test plans was the fact that the subject matter experts (SMEs) within each state who were responsible for the test specifications had very different views of the KSAs required to assure public protection. The results of Nelson's (1994) study underscores the crucial role of SME judgment in the development of any credentialing exam.

---

[1] Terminology notes: (a) The term credentialing is used here to refer to both licensure and certification—there is, however, a distinction between the two terms which is not addressed in this paper; (b) The acronym KSAs commonly refers to knowledge, skills, and abilities. However, in credentialing the emphasis is on knowledge and skills, not abilities; (c) the term occupations is occasionally used to refer to both occupations and professions; (d) the term practitioner refers to workers, incumbents, or others who practice an occupation or profession; (e) the terms test specifications and test plans are used interchangeably.

BEST COPY AVAILABLE

The purpose of this paper is to review and evaluate methods for conducting job analysis. The remainder of this section presents a general framework for conceptualizing the various approaches to job analysis. The second section addresses four critical research questions related to the practice of job analysis. The third and final section considers the degree to which job analysis can contribute to the protection of the public, examinees, and credentialing agencies

Dimensions of Job Analysis

A job analysis is a systematic procedure for collecting and analyzing job-related information to meet a particular purpose (Ash, 1988; Gael, 1983; Primoff & Fine, 1988). There are four key factors or dimensions along which job analysis methods vary (Ash, 1988; Cornelius, 1988; Levine, Thomas & Sistrunk, 1988; McCormick, 1976). Although there are slight differences of opinions regarding how to organize and label those four factors, the following are common to most lists: (1) the types of descriptors used to characterize the job; (2) the sources of job information; (3) data collection methods; (4) level of detail or level of analysis. Each of these four factors must be considered when making decision about the specific methods for a study.

- ***Types of Job Descriptors***. Job analysis involves reducing to words the things people do in work (Ash, 1988). Because language is rich, complicated, seemingly infinite, and ambiguous, condensing a job into words is a challenging endeavor. In order to adequately describe a job in a systematic manner, it is necessary to identify and use descriptors that are both accurate and have a common meaning to those who will use the results of a job analysis. One of the most significant distinctions among job analysis methods is whether they focus primarily on characteristics of the person or on characteristics of the job or task (e.g., Ash, 1988; Dunnette, 1976; Fine, 1986; Pearlman, 1980; Fleishman & Quaintance, 1984, p. 48, McCormick, 1976). In general, ***task-oriented*** procedures focus on the tasks and behaviors necessary to result in some output, product, or service. For example, the statement "transcribes letter using standard format," is a task-oriented descriptor of work behavior. In contrast, ***person-oriented*** approaches to job analysis focus on the human attributes required to successfully complete the tasks and behaviors required of a job. These human attributes would included the knowledge, skills, abilities, attitudes, or personality characteristics required to achieve some specified level of performance. ***Other types of job descriptors*** might include the types of tools or equipment, the job context or work environment, the products made, or services rendered (Ash, 1988; Levine, Thomas, & Sistrunk, 1988). These other types of descriptors have not received much attention in the job analysis literature. However, they have an important place in the development of credentialing exams.

- ***Sources of Information.*** Data about a job, occupation, or profession can be obtained from individuals familiar with the job or from existing documents and records. Many job analyses published in the personnel psychology literature make use of job data obtained from a wide variety of individuals, including actual practitioners, supervisors, managers, job analysts, and SMEs. Written documentation can also be helpful. Job descriptions, quality control analyses, insurance records, patient charts, and textbooks would all qualify as legitimate sources of information for a job analysis. Most job analyses will rely on multiple sources of information.

♦ ***Data Collection Methods.*** Job analysis data for a credentialing exam can be obtained using of a wide variety of procedures commonly used in the social sciences. The more common methods for eliciting and/or recording information include direct observation, interviews, job diaries, task inventory questionnaires, focus groups, and SME conferences. In addition to relying on one or more methods of data collection, conducting a job analysis requires that decisions be made regarding sampling. The sampling plan must consider multiple universes, including work settings, sources of information, and tasks. The sampling plan must also consider whether each universe is sampled randomly or systematically. Sampling plans may dictate a specific method, or visa versa. For example, the critical incident technique strategically samples only critical job behaviors.

♦ ***Level of Detail.*** Job analysis methods vary according to the level of abstraction of the descriptors used to depict a job. The box below presents pairs of job descriptors at different levels of abstraction. It is obvious that the use of one or the other of these pairs would influence the outcomes of a job analysis. Either specific or general descriptors could have a legitimate use, depending on the specific purpose of the job analysis. Level of detail is an important consideration in a job analysis and will influence other aspects of the study, such as data collection methods, source of information, and specificity of the test specifications.

LEVEL OF DETAIL	
Specific	General
balance accounts receivable ledger	number facility
explain structure of left ventricle	human anatomy and physiology
understand an apartment lease	written comprehension
conduct psychotherapy for patient with substance abuse disorder	counseling

Research Questions

Over the past two decades, numerous investigations of job analysis methods have been conducted to determine the extent to which variations along the four dimensions outlined above affect the outcomes. The next section of this paper addresses some critical issues related to the research and practice of job analysis within the context of licensure and certification testing. This paper does not attempt to provide a comprehensive review of the job analysis research to date. A concise summary of job analysis research is provided by Cornelius (1988). Other reviews can be found in journal articles and book chapters in personnel psychology (e.g., Dunnette, 1976; Fleishman, 1982; Fleishman & Quaintance, 1984; McCormick; 1976; Pearlman, 1980). Interested readers are referred to that body of literature. That research addresses topics such as job analysis as applied to job evaluation, the creation of job families to support validity generalization, the development of personnel selection tests, the reliability of task ratings, and other topics. An excellent reference on job analysis is the two volume set edited by Gael (1988). This text contains over 80 chapters on job analysis research and practice. It is interesting to note that not a single chapter addresses job analysis for licensure and certification testing.

When evaluating that literature, it is important to consider how job analysis for licensure and certification differs from job analysis in the general area of personnel psychology. One major difference is that job analysis for licensure and certification testing usually requires a broader sampling plan than job analysis conducted in business and industry. Job analysis projects for credentialing exams are typically national in scope, and may cover a multitude of settings. Since licensed or certified individuals, by way of their credential, are deemed qualified for a variety of positions, it is necessary for the job analysis to include multiple jobs. Another major difference is the singularity of purpose of a job analysis for a credentialing exam — the development of test specifications. Another difference is that credentialing exams measure the *knowledge and skills* required to function competently, and typically do not address *abilities and aptitudes*. Furthermore, most credentialing exams usually focus on acceptable levels of performance instead of outstanding performance. There are exceptions to this. The following text discusses four issues: (1) the effectiveness of common job analysis methods; (2) the optimal type of descriptor for characterizing a job; (3) the identification of KSAs; (4) translating job analysis results into test specifications.

QUESTION #1: *Which methods for conducting job analysis for credentialing exams are most effective ?*

The four dimensions outlined above suggest an almost endless supply of methods for carrying out a job analysis. In practice, however, a handful of approaches prevail. In the larger field of job analysis, some of the more common procedures include task inventories, the critical incident technique, the Job Diagnostic Survey, functional job analysis, the Position Analysis Questionnaire, and the Ability Requirements Scales. Some of these methods, as well as others, have been found suitable for use in developing credentialing examinations. The following text considers the task inventory method, the professional practice model, the critical incident technique, functional job analysis, and other methods.

Task Inventories

One of the most common methods for collecting and analyzing job-related information is the task inventory survey (Gael, 1983). A task inventory questionnaire consists of a list of job-related activities and behaviors thought to be important for a particular occupation. An initial list of tasks is developed based on direct observation, interviews with practitioners or SME panels, job diaries, and other sources of information. These tasks are then formatted into a survey and rated by practitioners in terms of certain characteristics, such as frequency, importance, time spent, criticality, or difficulty of learning (Kane, Kingsbury, Colton, & Estes, 1989; Sanchez & Levine, 1989; Sanchez & Fraser, 1992; Schmeiser & Estes, 1986). The goal is to identify important tasks, because the test specifications should emphasize such tasks. Task importance is often defined as a function of other, more specific task ratings. For example, Kane et al (1989) proposed a method for weighting tasks according to their ratings of task frequency and task criticality.

Numerous job analysis studies conducted within the context of licensure and certification testing have made use of the task inventory method (e.g., Fidler, 1988; Kane, Kingsbury, Colton & Estes, 1986; National Council of Examiners in Engineering and Surveying, 1989; Lunz, Stahl & James, 1989; Rosenfeld, Shimberg,

& Thornton, 1983). In addition to presenting long lists of tasks, some studies have asked respondents to rate KSA domains in terms of relevance or importance (American College Testing, 1991; Rosenfeld, et al, 1983). Other studies have focused exclusively on KSA domains while excluding task (Tannenbaum & Wesley, 1993).

The task inventory method has several advantages, which may explain its widespread use. First, since task inventories are relatively easy for respondents to complete in a short amount of time, it is possible to include a large sample of tasks on a survey. Second, task inventories provide a convenient vehicle for sampling hundreds of individuals in numerous work settings (Cascio, 1982; Kane et al., 1986). The breadth of coverage provided by a task inventory survey is especially important for licensure and certification examinations, which are intended to indicate readiness for a wide range of activities in a variety of settings (Kane, 1982). For example, radiologic technologists may work in physicians offices, rural hospitals, large urban hospitals, university hospitals, or podiatrists offices, to name a few settings. Third, data resulting from a task inventory survey are conducive to many types of statistical analyses (Cascio, 1982). Such analyses might be used to organize tasks into a meaningful model (factor analysis), or to compare profiles of practitioners in different settings or with different types of experience (ANOVA, profile analysis), or to create of job families (cluster analysis). Fourth, data from task inventories lend themselves to the development of test specifications based on empirical methods (Kane, et al., 1989; Raymond, 1994; Schmeiser & Estes, 1986). Finally, the task inventory method provides tangible numbers. Such data can be useful in convincing the courts that an exam is job related (Thompson & Thompson, 1982).

The task inventory is not without a few noteworthy limitations. For starters, a task inventory, as with any written survey, carries with it the liabilities of written language. No matter how carefully the task statements on survey are written, some will not be interpreted as intended. Sometimes, follow-up telephone calls are required to determine what respondents really meant. Second, the use of checklists and rating scales provides an incomplete picture of a job. Not only is the interdependence among tasks lost, but it is difficult to gain an overall perspective of a job based only on task inventory responses. Third, respondents may not always be qualified to make the judgments the survey is intended to elicit. For example, many task inventories ask respondents to provide criticality ratings, or other complex judgments. However, criticality is a complicated construct; not only is it multidimensional, but it often requires in-depth knowledge of difficult topics. Fourth, task inventories focus on the tasks people perform. Many professions depend on unobservable knowledge, cognitive skills, and professional judgment; an emphasis on observable tasks may overlook the cognitive nature of many professions. In addition, task inventories seem to emphasize the *verb* component of a job. This is not to suggest that current task inventories completely neglect nouns, but that there seems to be a preoccupation with verbs. A more useful focal point might be the *nouns* — the things that people work with (i.e., equipment, types of problems, coworkers, information, patients, etc.)[2].

[2] Footnotes are for speculation: People work with nouns, and these nouns influence what a person does on the job. Once the nouns are known, we can often make inferences about the verbs that occur. If we know that a counselor or psychiatric nurse works with psychiatric patients, do we need to know how many times a week they restrain patients, administer medications, conduct therapy sessions, attend meetings, and so on? Perhaps we would be better off knowing about the types of patients they deal with, and the general scope of practice (e.g., monitoring, diagnosis, treatment), and the methods/ modalitities they use in their practice (i.e., behavioral models, drug treatment).

Anyway, the task inventory is not the perfect approach to job analysis. Perhaps it is some of the limitations noted above that prompted LaDuca to contend that "task analysis is highly inappropriate with complex, high status professions, such as medicine or law" (1984, p.183), and to recommend an alternative based on a model of professional practice.

Professional Practice Model (PPM)

The professional practice model[3] is a general framework for conducting job analyses that has been used primarily in the health professions (LaDuca, Engle, & Risley, 1978; LaDuca, 1980; 1994; McGaghie, 1980). The PPM encourages a comprehensive description of an occupation, including job responsibilities, KSAs, and the context in which a job is performed. The job context might include social factors (other workers, the work setting, organization design), technological factors, or the types of information required to perform a job. The key feature of the PPM is that it is not task bound. The PPM seems well-suited to many occupations, and produces the types of information relevant for credentialing. In medicine, and many other occupations, the job analyst and test developer are really interested in: (a) the types of problems a practitioner encounters; (b) the human, material, and informational resources that should be used to solve those problems; and (c) the other that may facilitate or inhibit the solution of those problems.

An important feature of the PPM is that it is based on principles of facet design. The model requires its users to first identify and organize the major dimensions or facets of job performance. Once the underlying dimensions are defined, the categories of job performance within each dimension are specified. A simplified model of physician practice might consist of two facets: care setting and type of medical problem. The care setting facet might be comprised of two categories (hospital and physician office/clinic), while the medical problems might be organized according to anatomic systems (e.g., endocrine, skeletal, etc.). Other facets might include patient age, gender, or severity of the medical condition. The intersections or cells produced by crossing the facets give rise to numerous job-related situations. It is these situations that determine the job responsibilities and KSAs required of a practitioner. Another important characteristic of the PPM is the production of scripts — the identification of the KSAs required to successfully manage selected situations (LaDuca et al, 1984). These KSA statements form the basis of the test specifications. The PPM has been used to describe practice in occupations such as medicine, medical technology, occupational therapy, and medical dietetics (LaDuca, 1980; LaDuca, et al, 1978; LaDuca et al, 1984; McGaghie, 1980).

One advantage of the PPM is that it provides a top-down, model driven approach to conceptualizing the world of work. The model does not identify isolated, discrete tasks, but specifies the universe of problems that a practitioner may be called on to solve. For most professions, this is what we are really interested in. The PPM also identifies the situational factors that contribute to or inhibit the solutions to those problems. Once we understand how those factors and how they interact, we've got a pretty good handle on what a job entails. Another important advantage is that facet models can be used to specify work-related situations, which can then be used as the raw materials for test questions. One potential limitation of the PPM is that some of the studies have relied extensively on panels of SMEs, particularly for the specification of KSA requirements. This

[3] Called the professional performance situation model in some writings.

is a limitation of most other methods. One way to verify SME judgments would be to survey practitioners and other relevant parties regarding the relevance of the results produced by the SMEs. In fact, the PPM is open to various data collection methods. Another potential limitation is that the results will depend to some extent on the form of the model that is used at the start of the project. However, there certainly is nothing that says a model cannot be modified or abandoned in the midst of a project. It appears that in the study reported by LaDuca et al., (1984) the model was rather dynamic, changing direction as more information was obtained.

Flexibility seems to be a key attribute of the PPM. It is not really a specific method of job analysis, but a conceptual framework for guiding efforts to understand the entire context of what practitioners do and what they need to know to do it. The professional practice model is worthy of use in more job analysis work.

Critical Incident Technique (CIT)

The CIT was documented by Flanagan in 1954. Since that time, it has been used to study numerous jobs, occupations, and professions (Wilson-Pessano, 1988). The CIT describes jobs in terms of those behaviors that are expected to have a significant impact on the outcomes of work. This method involves the collection of anecdotes about job-related incidents that describe particularly effective or ineffective job performance. The anecdotes can be provided by practitioners, supervisors, or others familiar with the job. Each anecdote will typically include the following: the context in which the incident occurred and what lead up to the incident; what the individual worker did that was especially effective or ineffective; the consequences of the effective or ineffective behaviors; and other information that will help understand the incident or the context within which it occurred.

A CIT study is typically conducted by having qualified individuals complete a standard questionnaire, although interviews and focus groups can be used as well. A typical critical incident study can generate hundreds or even thousands of specific anecdotes (e.g., Jacobs, Fivars, Edwards, & Fitzpatrick, 1978). These anecdotes are then analyzed by SMEs and classified into more general job performance categories. Once this has been done, specific products such as test specifications, performance evaluation forms, and training curricula can be developed. The CIT has been used within the context of licensure and certification in several instances (Wilson-Pessano, 1988). A few of the more widely cited studies involved nursing (Jacobs, et al., 1978) general medicine (Hubbard, Levit, Schumacher, & Schnabel, 1965), and orthopedic surgery (Levine, McGuire, Miller, & Larson, 1968; Miller, 1968). Each of these studies generated thousands of critical incidents, which were subsequently classified into major domains of practice, which in turn, served as the basis for test content. Other studies have been conducted within the context of evaluating the performance of medical students and residents (Edwards, Currie, Wade, & Kaminski, 1993; Tarrico, Smith, Altmaier, Franken, & Van Velzen, 1984).

One advantage of the CIT is that it focuses on job requirements that are specifically related to outcomes (Kane, et al., 1986). This, in turn, should lead to test specifications with an obvious relationship to practice. A second advantage is that each critical incident can provide the raw materials for test questions. It is possible for item writers to create scenarios based on actual incidents, and then to ask a series of questions based on that scenario. This is a very useful procedure for generating practice related items. Another advantage is incidents can later be decomposed into component tasks, and included as part of a task inventory survey

(Rosenfeld & Thornton, 1976, cited in Cascio, 1982). There are also disadvantages to the CIT. The time and effort required to complete a study is one (Levine, Bennett, & Sistrunk, 1980; Cascio, 1982). However, a more substantive limitation is that the CIT, by definition, samples a very limited segment of the job domain — only those job behaviors believed to be most critical. This sampling strategy likely results in an incomplete picture of a job or occupation (Cascio, 1982). It could be argued that the CIT is not suitable for licensure or certification, where public protection is of primary concern. That is, the CIT may result in test specifications that overlook frequently performed, but only mildly critical behaviors. Mildly critical behaviors might be important to address if large segments of the public are exposed to such behaviors. A credentialing exam should balance criticality of outcome with probability of outcome, and it is questionable whether the CIT provides this balance.

Functional Job Analysis (FJA)

The functional job analysis has evolved from job analysis projects conducted during the past 50 years by the US Employment Service (Fine, 1986), and is attributed to the efforts of Sydney Fine. Job information generated by the FJA serves as the foundation for *Dictionary of Occupational Titles* published by the US Department of Labor. The FJA is another comprehensive approach to job analysis in that it addresses tasks performed, KSAs required to perform those tasks, as well as the situational and social factors that influence job performance. Perhaps the most popular product to evolve from FJA is the framework for describing jobs and tasks in terms of their orientation toward *people, data, or things*. Although FJA has not been used extensively in licensure in certification, it is presented here for a few reasons. First, it is undoubtedly one of the most influential methods of job analysis, given its association with the Dictionary of Occupational Titles. Second, it has been discussed in the educational measurement literature, specifically as it relates to performance assessment (Fine, 1986). Third, it is systematic, rigorous, and comprehensive; and many of its principles can be applied to job analysis within the context of licensure and certification. For example, D'Costa (1986) presents a role delineation methodology that borrows concepts from the FJA.

The FJA provides a highly structured format for writing task statements, with each statement containing the following elements: the action that performed (verb); the recipient of that action (object of the verb); the purpose or product of the action; the tools, and the resources or equipment required to complete the action; and whether the task is prescribed or left to the worker's discretion. Here is an example of a task statement written according to FJA guidelines:

> *Types standard form letter, including specified information from record provided, following standard operating procedure for form letter, but adjusting standard form as required for clarity and smoothness in order to prepare letter for mailing* (adapted from Fine, 1986, p. 63).

Once task statements are written, they are rated by SMEs in terms of the worker functions required to perform each task. The scales for accomplishing this (worker function scales) indicate the extent to which each task involves working with people, data, and things, and the exact nature of that involvement. For example, a task that requires working with data might require data synthesis, computation, or both. Each task is also rated according to the language, math, and reasoning skills that are required. Tasks are also evaluated in terms of

physical strength. Finally, each task is described in terms of any special KSAs required to perform those tasks, training requirements, any equipment that may be necessary, and the level of performance required in order for task performance to be judged acceptable. Most of the information for a functional job analysis is generated by SMEs.

One of the important contributions of FJA is its insistence on the use of a structured format for producing task statements and KSA requirements. The use of rules and principles to guide our use of language within the context of job analysis can only improve the results. An obvious advantage of FJA is that it is comprehensive. It addresses not only task, but the KSAs required to perform those tasks. As Fine has noted, "the issue is not whether to be work oriented or worker oriented. Job analysis must be both, and it must focus on the environment of the work organization as well." (Fine, 1986, p. 60). Another admirable quality of the FJA is its level of specificity. But, specificity is a two-edged sword. On one hand, the level of detail required of the FJA is suitable for studying a particular job. On the other hand, a great level of detail may not be feasible for studying entire occupations and professions. Consider applying the FJA to the nursing profession. Although many nurses perform the same tasks, the manner in which they perform them varies considerably from one setting to the next. It would be a tedious activity for SMEs to specify factors such as equipment requirements, social environment, and level of supervision for al settings in which nurses work. It would presumably take multiple SMEs from each of these settings to provide accurate information.

Functional job analysis as describe by Fine (1986) relies extensively on small panels of SMEs and trained job analysts, which limits its utility for licensure and certification programs that are national in scope. In business and industry, there is a strong tradition of using trained job analysts and small samples of SMEs, although such an approach has become less popular in recent years (Landy & Vasey, 1991). One way to overcome this limitation is to combine the FJA with that of the task inventory. Through such a combination, SMEs would provide the information that they are most qualified to provide, and larger samples of workers would provide additional information (e.g., task frequency) or would be used to support of challenge SME judgments. Another limitation is that many of the FJA scales (e.g., physical strength, language skill) would not be suitable for licensure and certification testing. As noted by Fine (1986), the scales, although practically useful, may lack empirical validity. A final limitation is that FJA requires a significant investment of time and energy to learn to use the FJA method. By Fine's (1986) estimation, a week of basic training and six months of supervised experience is required. Nonetheless, the FJA provides many sound ideas and principles that could be used by other job analysis methods.

Combined Methods

There is a growing consensus that a job analysis project based on multiple methods is likely to be more useful than a project based on any single procedure (Levine, Thomas, & Sistrunk, 1988). One notable example of a combined approach is the role delineation model advocated by D'Costa (1986). This method, like the FJA, conceptualizes important job functions in terms of their relationships to people, data, or things. D'Costa's role delineation method also emphasizes the social context within which a clinical problems occur, and describes practice in terms of the intersections of a multifaceted model. The ultimate goal of the role delineation method is to develop a comprehensive model of practice. Thus, it similar in some ways to the

professional practice model advocated by LaDuca and colleagues. The role delineation method discussed by D'Costa also makes use of a task inventory, and the judgments of SMEs to specify KSA requirements. And finally, this method also has something in common with the critical incident methodology. The facet model resulting from a role delineation study can be used to specify critical incidents or scenarios, which can then be used as the basis for writing practice-based test questions (D'Costa, 1986).

Other examples incorporating multiple methods can also be found in the literature. One such study was conducted for the purpose of validating and updating the test specifications for a certification examination in nurse anesthesia (ACT, 1992; Zaglaniczny, 1993). A panel of SMEs identified four facets that described most clinical situations encountered by a nurse anesthetist: (1) patient health status; (2) medical or surgical procedure; (3) anesthesia process (agents and technique); and (4) instrumentation required. For each of these facets, a list of elements belonging to that facet were identified and formatted into a survey. The individual elements were then rated in terms of frequency of occurrence and level of expertise required. Other sections of the survey considered additional facets (e.g., patient age; work setting). The final section asked respondents to assign weights to the knowledge domains that were being considered for inclusion on the test specification. The weights assigned by practitioners were later compared to the judgments of SMEs. A job analysis of psychologists (Rosenfeld, et al, 1983) also used a combination of job analysis models. Although that study used a survey, it was model-driven and did not employ a traditional task inventory survey. The survey included 59 job responsibilities (not specific tasks), a list of 62 procedures and techniques, and 49 knowledge areas.

So, which method works best? In this instance, the safe response is probably the honest one. Although no single method works best, all approaches have obvious strengths, and a job analysis that borrows ideas from all methods is likely to be effective. The methods described above do, however, raise many issues pertaining to the sampling of tasks, level of detail, the inclusion of KSAs on surveys, reliance on SME panels, and other matters. Issues concerning some of these points are discussed in subsequent text.

QUESTION #2: ***Should descriptors for characterizing occupations and professions be task oriented or worker oriented?***

Task-Oriented Descriptors

Many job analyses completed within the context of licensure and certification have focused primarily on tasks and/or job responsibilities. This makes sense — in order to determine the KSAs, it is first necessary to have a firm understanding of the tasks or responsibilities for which the KSAs are presumably required. There have been some exceptions. A recent job analysis of the teaching profession focused exclusively on KSAs (Tannenbaum & Wesley, 1993). Other studies have included both tasks and KSAs (Rosenfeld, et al., 1985; LaDuca, 1984; ACT, 1991; Zaglaniczny, 1993). Including KSAs as part of a job analysis is likely to generate much useful information, but it is important that tasks or job responsibilities be included as well. The importance of a task-based approach to job analysis has been emphasized by scholars and supported by the courts. (D'Costa, 1986; Kuehn, Stallings, & Holland, 1990; Levine, et al., 1988; Thompson & Thompson, 1982;). Any credentialing program should be able to demonstrate a link between the KSAs skills tested by the exam and the tasks performed on the job. Documentation of the tasks or job responsibilities can only support

one's efforts to demonstrate such a link. Thus, it is important for a job analysis to identify what people do. This can be accomplished by the various job analysis methods previously described.

Person-Oriented Descriptors (KSAs)

The primary limitation of the task-based approach is that most credentialing examinations do not directly evaluate a candidate's ability to perform tasks. Most credentialing exams emphasize the cognitive KSAs that provide the foundation for competent task performance — these are the type of skills that can be conveniently and accurately assessed with multiple-choice tests (Kane, 1982; 1986; Schafer, Raymond & White, 1992). Since these KSAs are ultimately translated into the test questions that comprise an examination, it is only fitting that a job analysis, at some point, explicitly identify the KSAs required to function in a safe and effective manner.

One reason why KSAs should be explicitly stated is to assure that the credentialing exam specifies only relevant KSAs at an appropriate level of performance. Consider a set of test specifications comprised only of tasks. One of those tasks might be "*Assess patient's tolerance for physical activity.*" This task might appear on a task inventory for fitness trainers, nurses, occupational therapists, MRI technologists, or other health professionals. A few of the cognitive objectives that such a task could generate are listed below.

- determine the types of physical activities that would benefit a particular patient.
- differentiate normal from abnormal blood pressure results.
- identify the proper steps for assessing tolerance for physical activity.
- demonstrate knowledge of how various assessment instruments function.
- demonstrate knowledge of cardiovascular anatomy and physiology.

These objectives cover much ground. And yet, given the task statement presented above, it would be possible to justify the inclusion of test questions addressing any of these cognitive objectives. Reducing the list to just one of these objectives still leaves much room for variability. Here are just a few questions that address cardiovascular anatomy and physiology as related to the task, "*Assess patient's tolerance for physical activity.*"

- Which of the following heart rates might be regarded as abnormal?
- What blood pressure measurements fall within the normal range?
- Which of the following empties blood into the right atrium?
- What chamber of the heart has the thickest layer of myocardium?
- Why does the left ventricle have a thicker myocardium than the right ventricle?

Without a doubt, the level of knowledge required to answer these questions varies considerably. The problem is to determine the appropriate level of knowledge required to guarantee a specified level competence. Fortunately, the SMEs who serve on examination committees will prevent too many out-of-level test questions from finding their ways onto credentialing examinations. However, the risk is present unless the test specifications are clear. Test specifications that specify only tasks are insufficient for informing item writers, candidates, test assemblers, and exam committees of the test content. Specifying only tasks identifies neither the breadth of knowledge nor the depth of knowledge required.

Another reason why a job analysis and/or the test specifications should include reference to KSAs is because the results based on tasks are not always useful. The task "*verify the identity of the patient,*" or variations of it have appeared on many job analysis surveys. More often than not, this is one of the most frequently performed and critical tasks for a wide variety of professions. No one disputes this. However, the importance of this task also implies that it should receive considerable weight in test specifications, at least when the test specifications are empirically derived. (e.g., Kane, et al, 1989; Lunz et al, 1989; Raymond, 1994). This is where the problem lies. There is a limited number of questions that can be asked about this task, because there is a very limited knowledge base associated with this task. Should such a knowledge–impoverished task receive substantial weight on a set of test specifications? Probably not. A related problem is that examinees, item writers and others will ask, *"OK what do I need to know about verifying a patient's identity?"* For this particular task, it is probably obvious, but for other tasks, it will not be so obvious, as argued in preceding paragraphs. For these and other reasons, a job analysis project should address both tasks and KSAs. The practice of deigning a study to include both tasks and KSAs is becoming more common (Ash, 1988; D'Costa, 1986; Fine, 1986; Hughes & Prien, 1989; Landy, 1988; Levine et al, 1988).

QUESTION #3: ***How should KSAs be specified ?***

A job analysis might be viewed as consisting of two distinct, but interdependent components: (1) the identification of job responsibilities and (2) the specification of KSAs required to effectively perform those job responsibilities. Although both of these components are important, the following text addresses only KSA domains. Specifying KSAs is a particularly complex endeavor dealing almost entirely with abstractions, and obtaining valid information is no easy feat. Since most credentialing examinations address knowledge and cognitive skills, specifying KSA domains is obviously a critical activity. The discussion considers three methods for specifying KSAs: surveying practitioners; establishing empirical links between tasks and KSAs; and asking SMEs to make judgments regarding KSAs.

Surveying Practitioners about KSAs

Several job analysis surveys in recent years have focused primarily on, or exclusively, on KSAs. (Tannenbaum & Wesley, 1993). A KSA survey might consist 100 or more phrases and respondents are asked to rate each KSA in terms of importance, relevance, or some other construct. One advantage of this approach is that KSAs which eventually comprise the test specifications are based on the responses of large representative samples of practitioners. It would seem that potential litigants would have a difficult time disputing test specifications based on a representative sample of peers. A second advantage is that KSA ratings can be useful for establishing weights for test specifications. A third advantage of including KSAs on job analysis surveys is that complex professions are characterized not only by tasks and job responsibilities, but also by the KSAs that professionals bring to bear on those tasks (LaDuca, et al., 1984; LaDuca, 1994). Furthermore, it is possible that individuals working in a profession who perform the same tasks may require different constellations of KSAs, and including KSAs on a survey could provide the means for discovering such differences. Consider the task "*provide dietary counseling to a patient receiving chemotherapy.*" The KSAs

required to perform this task with a 10-year old patient would be different from the KSAs required to perform the task with 70-year old patient. A fourth advantage of including KSAs on a job analysis survey is that, if tasks are also included on the same survey, then it may be possible to establish empirical linkage between tasks and KSAs by using multivariate statistical procedures (Rosenfeld, et al, 1983).

There are, however, a number of serious limitations to obtaining judgments about KSAs on job analysis surveys. The first problem pertains to the difficulty of defining individual KSAs. Unlike tasks, KSAs represent abstract, complex, and potentially ambiguous concepts. Although the ambiguity can be eliminated with clear definitions, it is challenging to arrive at a concise and workable definition within the confines of a survey. As an example, the 50 or so abilities that comprise the Ability Requirement Scales (Fleishman & Quaintance, 1984) took years to identify and define. Although the definitions are clear, they are hardly brief, often consisting of 30 to 40 words. Including good definitions on a survey of 100 to 150 KSAs would surely result in a dense and complex instrument.

Are definitions really necessary? Consider the KSAs in the box below. All were taken from recent job analyses. The first KSA in the list (*quality assurance*) received the highest importance ratings of 35 topics included on a survey for managers of medical laboratories. Although this KSA seems straightforward, it does beg a few questions. For example: (a) does *quality assurance* include quality control; if so, does it include all of the QC tests related to specific outcomes, as well as generic QC methods such as statistical process control? (b) does *quality assurance* include total quality management and embrace topics such as Deming's 14 points or the development of outcome indicators? (c) does quality assurance include topics like the use of Pareto charts (a topic in many QA manuals)? Such questions may arise in the minds of survey respondents. These questions point to a single issue: that of establishing boundaries for each KSA and clarifying breadth of coverage. This can be accomplished by providing definitions or by presenting examples. Granted, there may instance for which definitions really aren't necessary. For example, if *quality assurance* was followed by *total quality management*, the presence of the latter would help define the former. It is also possible that the meaning of *quality assurance* as it applies to laboratory management would be obvious to all survey respondents, thereby obviating the need for a definition. Such would be the case if, for example, *quality assurance* were part of a core curriculum, and all educational programs covered the same topics related to *quality assurance*.

EXAMPLES OF KSAs FROM SURVEYS

KSA domain	Occupation
operations management: quality assurance	laboratory manager
probability & statistics	engineer
counseling and guidance techniques	psychologist
knowledge of where to obtain information	Insurance agent
computer operations: data storage and retrieval	MRI technologist
reading skill: recognize or predict ideas or situations	teacher

As another example, consider the final task in the list (*reading skill: recognize or predict ideas or situations*). It appeared on a job analysis survey for a basic skills test for teachers. This KSA has something to do with reading comprehension, although it is difficult to say exactly what. Part of the problem is due to the conjunctive nature of the statement ("or" occurs twice thereby creating up to four separate KSAs). However, much of the ambiguity is due to the words "ideas" and "situations." Survey respondents will likely have different interpretations of this KSA, which will limit the validity of the ratings. As it turns out, this seemingly fundamental KSA was flagged for possible exclusion from the exam, due to its relatively low importance ratings. Perhaps the low ratings were a symptom of its ambiguity.

A second area of concern with the inclusion of KSAs on surveys pertains to the rating scales and the quality of the ratings that are obtained. The validity of ratings will depend on factors such as how well each KSA is defined, the number of scale points, the types of scale anchors, and the ability of the respondents to understand and use the rating scales. Many recent job analyses have obtained importance ratings for KSAs by using a Likert-type scale. Although this seems to make good sense, KSA importance may mean different things to different people. It could mean frequency, criticality, depth of knowledge, time spent, or level of proficiency required. Even when scales are apparently well designed, rating errors are difficult to eliminate. One type of error is halo – the tendency for respondents to attend to one salient feature when making their ratings. A casual review of the literature will reveal instance where scales measuring somewhat independent constructs are more highly correlated than one would expect (e.g., .60s to .80s), which leads one to question just what the two scales are measuring. Another potential problem with KSA scales is that they are susceptible to a positive response bias (Landy, 1988, p. 277). Uniformly high ratings can make it difficult to differentiate between the truly important and only moderately important KSAs.

Further research is needed to determine the KSA attributes that are most worthy of measurement, and to develop scales that measure those attributes. The use of importance scales may not result in the type of information required to assure the validity of a credentialing exam. As noted by Kane (1982), "the level of proficiency required by the examination should not be higher than that required for practice," implying (to me) that KSAs should be judged in terms of level of proficiency or depth of knowledge. Scales based on Bloom's taxonomy or some other theoretical framework might be useful for rating knowledge domains, or at least might provide a frame of reference for additional work on scale development. An example of level of knowledge scales corresponding to Bloom's taxonomy can be found in the job analysis of licensed psychologists (Rosenfeld, et al., 1985). Other models of cognitive functioning might also be of use. The rating scales for the Ability Requirement Scales (Fleishman, 1975; Fleishman, 1982; Fleishman & Quaintance, 1984) provide an excellent model for the development of scales to measure proficiency, level of knowledge, and related constructs. Each of the Ability Requirements Scales is preceded by a detailed definition and a description of how that ability differs from other, similar abilities. Then, a 7-point rating scale is presented, with points on the scales anchored to specific behaviors. Because these scales correspond to abilities, rather than knowledge or skills, they have limited applicability to job analysis for credentialing exams. However, the ideas and procedures for developing the Ability Requirements Scales are worthy of replication.

Another limitation of identifying KSAs using large-scale surveys is that the questions being asked of survey respondents do not make reference to specific tasks. That is, respondents are asked something like, "How important is this KSA to your job?" Due to the complexity of most occupations, it is probably unreasonable to expect survey respondents to simultaneously consider all tasks and then make an overall judgment concerning the KSAs required to complete those tasks (Landy, 1988). This problem can be addressed in part by including both tasks and KSAs on a survey, thereby giving respondents an idea of the tasks they should consider. However, this is less than optimal. Several sources note the importance of establishing a clear link between job responsibilities and the KSAs to be tested on an examination (AERA, et al., 1985; D'Costa, 1986; Levine, Thomas, & Sistrunk, 1988; Primoff & Fine, 1988; Thompson & Thompson, 1982). Thus, the real question that survey respondents should be addressing is something like "*what level of proficiency is required for this KSA in order to perform task X*?" Obviously, it would not be feasible to have respondents rate each task in terms of every KSA on a mail-out survey.

Most licensed occupations are characterized by a complex network of tasks and KSAs. Identifying the components of this network and the relationships among those components is no trivial undertaking. KSA domains are often cast in a language that may be truly familiar only to educators, psychologists, or scholars in a given occupation. Furthermore, the individuals who provide KSA judgments should participate in some type of training in order to provide valid information (Fleishman & Quaintance, 1984, chapter 13; Landy, 1988). Therefore, another problem with including KSAs on job analysis surveys pertains to the qualifications of those who provide the information. Although typical practitioners are the best source of information about the tasks and job responsibilities they perform, the typical practitioner may not have the background and experience required to make valid judgments about KSAs, particularly if a mail-out survey is used. Judgments about KSAs should be obtained from SMEs.

Empirical Methods for Linking Tasks and KSAs

One of the most interesting efforts to identify and link KSAs to job responsibilities is described in the job analysis of licensed psychologists conducted by ETS (Rosenfeld, et al, 1983). The first section of the job analysis survey consisted of 59 job responsibilities thought to be important for psychologists. Another section of the survey consisted of 111 methods, techniques, and knowledge areas relevant to the practice of psychology. The process of empirically linking the KSAs to the job responsibilities required two independent analyses. First, the importance ratings for job responsibilities were factor analyzed, and four major categories of job responsibility emerged from the factor analysis. Then, a factor score was computed for each respondent on each of the four job factors. Next, respondents' ratings for the KSAs were correlated with the job factor scores. Correlations that exceeded .20 were taken as evidence of a linkage. In theory, this method is capable of detecting important differences in the way that different groups of individuals use KSAs in the work setting. For example, industrial psychologists may demonstrate a strong linkage between data collection techniques and factor 1 (research and measurement), whereas a school psychologists might show a strong link between data collection techniques and factor 4 (assessment).

Although this empirical procedure is certainly appealing, it might be regarded as only moderately successful. The correlations between tasks and job factors were generally low (.20 to .40), so that even the

strongest links were still tenuous. Limitations to this approach stem from the problems of including KSAs on surveys (discussed earlier) and problems related factor analyzing task importance ratings (discussed later). These criticisms notwithstanding, the method is fascinating, and that particular job analysis project serves as an excellent model for the credentialing community.

SME Panels for Judging KSAs

♦ ***Method 1: Panel Discussion.*** Relying on SME judgment is probably the most common approach to identifying KSA domains. Reasonably detailed explanations of how SMEs accomplish this are provided by Fine, (1986), LaDuca et al. (1984) and McGaghie (1980), and Raymond (1994). Detailed explanations are also likely to be found deep within the lines of various unpublished technical reports. Judgments from SMEs can be obtained in a very methodical fashion or in a relatively casual way. A reasonably methodical process might go like this:

1. SMEs meet to review the results of a job analysis and to identify important tasks based on those results.
2. Considering the important tasks, each SME prepares a list of KSAs. SMEs are asked to organize the lists into major categories. Sources of information for SMEs might include personal opinions, textbooks, articles, curriculum materials, other test content outlines, government reports, and so on.
3. The SMEs agree on major KSA categories and then list specific KSAs within each category. This is often done by having a recorder write KSAs on flip charts and then posting individual pages to the wall of the meeting room.
4. SMEs discuss, revise, and combine KSAs; KSAs are evaluated in terms of job relevance. Once consensus is reached, the meeting is adjourned.
5. The list is formatted into a document and mailed to SMEs and/or others for review and comment.
6. SMEs meet again to finalize the document and indicate degree of emphasis (i.e., weights) for each KSA.

This process will likely result in legitimate listing of KSA requirements. However, there are certain limitations to this approach. One limitation is that the KSAs are formulated with respect to the occupation as a whole rather than with respect to each specific task. Without referencing individual tasks, it is possible that the final list of KSAs will be deficient or will include potentially irrelevant KSAs. For example, SMEs may wish to include certain KSAs that are taught by most training programs even though the KSAs are not required in order to assure public safety. A related problem is that once a KSA makes it onto a list, it is often difficult to get rid of it. At best, a compromise is reached — the questionable KSAs get combined with or subsumed under other KSAs. Another concern is that this approach does not require SMEs to specify the depth of understanding or level of proficiency required for each KSA. In spite of these limitations, test specifications produced as outlined above are probably technically sound.

♦ ***Method 2: Task – KSA Linkage Activity.*** The process outlined above can be easily modified to address some of the limitations. After SMEs have listed all KSAs, they should participate in a "linkage activity." The purpose of the linkage activity is to establish an explicit association between each KSA and one or more tasks from the job analysis. For each task, SMEs are asked to identify the KSAs required to

successfully perform that task. The judgments could be dichotomous (yes, no) or could consist of ratings on a Likert-type scale (e.g., relevance, level of proficiency). This process requires that each KSA be independently justified. An additional benefit of the linkage activity is that the linkage ratings can be combined with ratings of task importance to produce weights for test specifications (Raymond, 1994).

One practical problem with the linkage activity is that it becomes unmanageable as the number of tasks and KSAs increases. A typical job analysis might result in a list of 150 important tasks and many, many KSA domains. In order to make the linkage activity feasible, the tasks and/or KSAs can be classified into broader categories (Landy, 1988). It is obvious that the validity of the linkage activity would suffer if the tasks and KSAs were condensed to only a few categories.

The linkage approach to establishing KSA requirements is advocated in the literature on personnel selection (Hughes & Prien, 1989; Landy, 1988; Fleishman & Quaintance, 1984). Landy (1988) demonstrates an application of this linkage method using the abilities comprising the ability requirements scales. Police officers were asked to indicate the extent to which each of 37 abilities was required to perform the tasks required of patrol officers. Rather than including each individual task, they were classified into 15 task categories. For each of the 15 task categories, police officers allocated points to indicate the extent to which each ability was required to perform that category of tasks. The points were constrained to sum to 100% for each task category. This rating method (constraining judgments to sum to 100%) may reduce the tendency to rate everything as important.

This type of linkage activity appears to be a sound approach to establishing KSAs for test specifications. However, its success depends on the use of clearly defined KSAs. Care must be taken to produce reasonably independent KSAs written at about the same level of specificity. Another pitfall SMEs should avoid is the tendency to add the phrase "knowledge of" or "skill at" to various tasks (Landy, 1988). Examples of this include, "knows importance of eliciting description of signs/symptoms in patients own words," or "skill at leadership." On the surface, such statements appear to be valid KSAs. However, these descriptors are broad and ambiguous, and really are not very helpful in supporting the test development process (Landy, 1988). Within the general field of personnel selection, the challenge of defining KSAs has been met by the development of taxonomies of abilities (Fleishman & Quaintance, 1984) and the affective traits related to job performance (e.g., Peterson & Bownas, 1984). The scales resulting from Fleishman's work are considered by some to be the most effective method to date for determining the abilities to included on personnel selection tests (Landy, 1988). Because licensure and certification generally is not concerned with abilities (AERA, et al., 1985), ability and trait taxonomies have limited applicability to credentialing. Could similar taxonomies be developed that would apply to the development of credentialing exams? I am skeptical for two reasons. First, abilities are relatively stable and finite, and therefore lend themselves to definition and classification. However, the knowledge and skills required for various occupations are virtually limitless and dynamic, and therefore will continue to elude attempts to define, parse and classify. Second, since numerous occupations exist, many such taxonomies are likely to exist. We may not need a separate one for each occupation, but one large, comprehensive taxonomy is not feasible and probably would not be useful.

Such pessimism is hardly constructive. What can be done at this point in time is to more carefully define the knowledge and skill constructs that SMEs work with, paying particular attention to issues such as level of

specificity and overlap. Second, additional work is needed to identify appropriate scales for rating KSAs. Scales based on frequency of use, relevance, level of proficiency, and depth of knowledge are possibilities. Concrete, behavioral anchors would be essential for these scales. Third, SMEs should receive adequate training in the use of scales for rating KSAs, and must be given ample time to use them. The task of linking KSAs to tasks is time consuming and tedious. But, it is essential.

QUESTION #4: ***How should the results of a job analysis be translated into test specifications?***

The Challenge

The ultimate purpose of job analysis is to develop test specifications that articulate the important characteristics of the test (e.g., content, number of items, cognitive processes, item formats, etc.). Although test specifications are often viewed as assisting those who develop tests, they have other functions are well. Some of the more important functions have been alluded to earlier in this paper and are summarized in the box below.

FUNCTIONS OF TEST SPECIFICATIONS

Provide direction to item writers.
Guide SMEs in reviewing test items and assembling tests.
Assure continuity in test content and difficulty.
Inform examinees of test content.
Provide a framework for equating, creating subscores, and doing other analyses.
Serve as legal documentation supporting the validity of test score interpretations.
Document the history of a certification program.

One of the fundamental challenges associated with the development of credentialing exams is that of translating the results of the job analysis into test specifications (Hughes & Prien, 1989; Kane et al., 1989; Schmeiser & Estes, 1986). The process for developing test specifications from job analysis data is complex, and requires extensive input from SMEs. First, SMEs are required to interpret the data produced by a job analysis. Such data would include the demographic characteristics of the sample, information about the work setting, as well as summary statistics for multiple measures for each specific task (e.g., frequency, criticality, time spent). Next, SMEs will need to specify the KSAs required to perform the important tasks identified through the job analysis. These judgments will be influenced by SMEs' understanding of the purpose of a credentialing exam, and differences in this fundamental understanding will surely impact the content of test specifications (Nelson, 1994). Finally, after considering all of this information, SMEs need to organize the KSAs into a coherent outline and assign weights to indicate the number of questions that should be allocated to each section of the test outline.

The influence of SMEs is illustrated in a study originally conducted to compare the effectiveness of different approaches to job analysis. Levine, Ash & Bennett (1980) evaluated the effectiveness of four job analysis methods using a variety of criteria, including the adequacy of test specifications eventually produced

from the job analyses. Levine et al. found notable differences among the four job analysis methods in terms of cost, utility of the data resulting from each method, and other factors. However, the four methods of job analysis resulted in four sets of test specifications that were very similar in content and quality. The authors concluded that test specifications "may be more a function of the insights and creativity of the person who must make the inferential leaps from the job analysis data to exams ... than of the quality of the job analysis data" (Levine et al., p. 534). Although it is encouraging that competent SMEs can compensate for low-quality job analysis data, it is somewhat disconcerting to concede that the test specifications are more sensitive to the "inferential leap" rather than to the actual data. This is a dubious state of affairs if we believe that testing procedures should have a clear and explicit link to the results of the job analysis.

There are really two ill-defined problems that must be solved when developing test specifications. The first is to structure the KSAs into some meaningful organizational framework. The second challenge is to assign weights to each KSA domain to indicate the degree of emphasis that each KSA will receive on the examination. These two issues are considered next.

Organizational Frameworks

A job analysis produces lengthy lists of tasks and KSAs, which must ultimately be translated into test specifications organized so that similar tasks or KSAs appear in the same categories, and different tasks or KSAs appear in different categories. Test specifications must be carefully worded and well organized if they are to serve their intended functions. Category labels such as "treatment planning" or "burn emergencies" should inform various audiences of the content of the test items that fall within these categories. If category labels do not serve these functions, then they probably are not valid. The process of imposing structure on test specifications is essentially a taxonomic activity (Schafer, Raymond, & White, 1992). The following text considers four methods for structuring test specifications: SME judgment; analysis of task importance ratings, analysis of similarity judgments, and analysis of task attribute ratings.

♦ ***SME judgment***. One method for organizing test specifications is to invite SMEs to discuss their ideas and arrive at some agreed-upon outline. SME judgment is most feasible when the domain is well understood, and when there is a general consensus concerning the manner for parsing that profession. One structured approach based on SME judgment has been used in conjunction with the critical incident technique (Jacobs, Fivars, Edwards, & Fitzpatrick, 1978; Levine, McGuire, Miller, and Larson, 1968). After the relevant incidents are identified, SMEs classify the incidents into categories based on their similarities and differences. This process for classifying critical incident is generally inductive and iterative. Although the judgments of SMEs usually result in sensible organizational frameworks that the public readily accepts, this approach is limited in at least two respects. First, because they are based on the opinions of a few SMEs, judgmental frameworks may not be replicable. This point was acknowledged by the authors of the critical incident study of orthopedic surgery (Levine, McGuire, Miller, and Larson, 1968). A second limitation is that judgmental frameworks are often based on the prevailing wisdom of the time, whether or not that wisdom has been supported by empirical research. For example, although Bloom's taxonomy has proven to be an extremely useful heuristic device for test developers and educators, scores based on collections of test questions classified according to Bloom's

taxonomy lack empirical validity (Cizek, Webb, & Kalohn, 1995; Kunen, Cohen, & Solman, 1981; Seddon, 1978). The limitations associated with SME judgments suggest the use of statistical methods.

♦ ***Multivariate Analysis of Task Frequency or Importance Ratings***. Organizational frameworks are also created using methods such as factor analysis or cluster analysis. One obvious advantage of these procedures is that categories can be defined in terms of some objective function. In addition, the procedures can be subjected to replication. One of the most common quantitative methods for structuring task and KSA domains is to factor analyze ratings from job analysis surveys, such as ratings of task frequency or task importance. In such instances, the original data matrix consists of ratings made by *n* persons on *p* tasks, and the matrix to be factored consists of correlations among the *p* tasks. This approach has come under increased scrutiny in recent years (Cranny & Doherty, 1988; Schafer, Raymond, & White, 1992). Most factor analyses of task ratings seem to result in *numerous factors with very few high loadings, accounting for small proportions of the variance*. Furthermore, the results are not always easy to interpret. The problem is that a useful factor analysis depends on variation and covariation — patterns of agreement and disagreement in ratings of task frequency among survey respondents. If subgroup X performs tasks 1, 3, and 5 while subgroup Y performs tasks 2, 4, and 6, then two clear factors will emerge: one corresponding to the behaviors of group X and the other corresponding to group Y. More often than not, the factor structure will reveal more about the people who provided the ratings (and their work environments), than about the similarity and differences among tasks (Cranny & Doherty, 1988). The problem is not really with factor analysis, but with the ratings to which the factor analysis is applied — ratings of task frequency, task importance, time spent, and so on (Schafer, Raymond & White, 1992). If we really want to know about the similarity and differences among tasks, then we should probably either: (1) ask SMEs for judgments of task similarity or (2) measure tasks on numerous attributes and cluster the tasks based on their measurements on those attributes. These alternatives are considered next.

♦ ***Multivariate Analysis of SME Similarity Ratings***. During the past 25 years or so, several studies have addressed taxonomic problems by applying multidimensional scaling analysis or cluster analysis to SME judgments of similarity. Similarity judgments can be obtained in a variety of ways. Schafer, Raymond and White (1992) asked 21 SMEs to sort 125 emergency nursing tasks into categories based on the perceived similarity of those tasks. A similarity index was computed for each pair of tasks, with similarity defined as the proportion of times that two tasks were placed into the same category by the SMEs. So, if 15 of 21 judges placed tasks A and B into the same category, the similarity index would be 15/21 = .71. The 125 x 125 similarity matrix was then transformed to distances and subjected to cluster analysis. The organizational framework resulting from this method was more interpretable and exhibited better empirical properties than the framework produced by clustering the frequency ratings provided by a national sample of emergency nurses. Related studies have been conducted in nursing (Raymond, 1989). orthopedic surgery (Raymond, Buckwalter & Steyers, 1989) and other fields (Rodgers, 1988; Sireci & Geisinger, 1992). A research team at the University of Iowa College of Nursing developed a comprehensive taxonomy consisting of hundreds of nursing interventions (Iowa Intervention Project, 1993; McCloskey & Bulachek, 1992). The taxonomy was formulated, in part, from similarity judgments subjected to multivariate analyses. Although these studies have used different measures

of similarity, in all cases similarity was the focus, not task frequency or task importance. Related procedures could make use of judgments concerning the relevance of specific test questions to each of several KSA domains (Messick, 1989, p. 39).

The use of SME judgments used in conjunction with multivariate grouping strategies shows promise as a mean for structuring performance domains. One advantage is that similarity is what we are trying to discover; thus, it makes sense to collect data that pertain to similarity. Second, such an approach combines the best features of the judgmental and empirical methods: it requires judgments from SMEs, and then subjects those judgments to algorithms that systematically derive categories based on well-founded principles. Third, the process can be subjected to replications with different tasks and different SMEs. Fourth, the results of most published studies have generally been quite interpretable by SMEs and testing experts.

♦ ***Multivariate Analysis of Measurements of Task Attributes.*** Preceding text has argued that applying multivariate analysis to ratings of task importance ratings (or frequency ratings) is not very beneficial. With such an analysis tasks are the variables, persons are the objects being measured, and the attribute being measured is something like "this is how important this task is to me." There are, however, legitimate applications of multivariate analysis to ratings of tasks, so long as the tasks are the objects being measured. It would seem quite reasonable to measure tasks on several relevant attributes, and then group the tasks according to their similarities and differences on the attributes. Such an approach would be consistent with classification methods employed in a wide variety of disciplines (Sokal, 1974). Fleishman & colleagues have developed several scales thought to be useful for measuring tasks, referred to as the *task characteristics rating scales* (Fleishman, 1982; Fleishman & Quaintance, 1984, chapter 13). The task characteristics rating scales are intended to evaluate tasks on attributes such as number of output units, precision of responses, stimulus duration, simultaneity of responses, number of steps, and numerous other attributes. Once tasks are measured on such attributes, it is relatively straightforward to cluster the tasks according to their similarities and differences on the measurements. Most of the research on the task characteristics rating scales have been conducted for jobs that involve machines and technical equipment (e.g., operating sonar equipment). I suspect that the task characteristics rating scales would <u>not</u> be helpful in developing test specifications for most credentialing examinations; however the methods underlying the development of those scales is worthy of consideration.

An interesting example of factor analyzing task attributes can be found in a widely-cited monograph on the Position Analysis Questionnaire ([PAQ] McCormick, Jeanneret & Mecham, 1972). The PAQ consists of 189 job elements; these elements correspond to a wide variety of job functions required in many occupations. One part of the study consisted of factor analyzing traditional job analysis ratings obtained from large samples of respondents in numerous jobs. The factor analysis resulted in five factors accounting for only 41% of the variance. A second part of the study obtained "attribute ratings" for each of the 189 elements. The attributes corresponded to 67 abilities and personal characteristics (e.g., perceptual speed, verbal comprehension, stage presence). To obtain the attribute ratings, personnel psychologists (SMEs) used a 6-point scale to rate each of the 67 attributes in terms of its relevance for each of the 189 job elements. The matrix of correlations among elements was then factor analyzed, resulting in six factors which accounted for 89% of the variance. This is a

pretty good fit. As noted by the authors, the results based on the attribute ratings were much more interpretable than the results based on the actual task ratings (McCormick, Jeanneret & Mecham, 1972, p. 357). There are two lessons in this study. The first reiterates the point that factor analyzing importance ratings is of questionable use. The second lesson concerns the mechanism suggested by the study for structuring test specifications. That is, it may prove fruitful to rate tasks in terms of the extent to which each task depends on various KSAs. The matrix resulting from such an exercise — the linkage matrix discussed earlier — can then be subjected to an appropriate multivariate grouping methods.

♦ ***Evaluating the Organization of Test specifications***. In many instances, SME judgments based on prevailing models of practice result in an effective test outline that is readily comprehended by the intended audience. In some instances, however, it may be prudent to rely on statistical methods when organizing test content. Applying multivariate grouping methods to judgments of task or KSA similarity can be very useful, but the traditional approach of factor analyzing task importance ratings (or frequency, criticality, etc.) is likely to be of limited use. One other approach that was not discussed earlier is to factor analyze item responses. For a variety of reasons, this method is not very useful for discovering organizational frameworks. The ultimate test of test specifications is how well they work in practice. Test specifications must have practical utility and empirical meaning.

- The category labels should be informative. Item writers should know what to write, test assemblers must be able to tell what an item tests by its content code, and examinees should be able to determine what to study. It should be possible for SMEs to accurately classify test items into content categories. Examinees who fail a test should be able to use their section scores to guide remediation efforts.
- Section scores should have empirical validity. This is essential for equating, and helpful for reporting and interpreting section scores. If correlations between section scores are near unity, and if score profiles are nearly flat, then the test specifications may be poorly organized. Sections should also be evaluated in terms of internal consistency (e.g., the internal consistency of a section can be compared to the internal consistency of a random collection of items).
- If there is doubt concerning the most effective framework, then competing organizational frameworks can be evaluated using both judgmental and empirical procedures.

The quality of test specifications will surely take on greater importance as computer-adaptive testing and automated test assembly become more common. Test specifications must have excellent organizational properties and sufficient detail in order for computer-based item selection algorithms to function effectively. It is interesting to note that the primary limitation of computer generated exams is that the computer often does not have adequate information about each test item in order to make intelligent item selections. Such limitations have been noted even with items coded according to very sound test plans (Swanson, Stocking, & Pearlman, 1993).

Establishing Weights for Test Specifications

Once the tasks and KSAs from a job analysis have been cast into an outline, matrix, or some other form, it is necessary to assign weights to test specifications. One approach for establishing weights is to ask SMEs

provide direct, holistic judgments concerning the number of questions to be allocated to the various sections of test specifications (Hughes & Prien, 1989; Raymond, 1994). One feature of holistic ratings is that they are made with respect to major categories of KSAs or major categories of tasks. For example, holistic weights for an examination in anatomy might be obtained simply by asking a panel of SMEs to assign percentages to each of several categories such as cardiovascular system, skeletal system, and so on, so that the percentages summed to 100%. SME judgments might be influenced by various types of information, such as the complexity of the anatomical system; the incidence or severity of pathological processes that invade each system, the amount of time devoted to that system in the curriculum, and the results of a job analysis. Another feature of holistic judgments is that, although SMEs might be asked to consider certain types of information that may be available to them, specific algorithms for evaluating and integrating that information are not imposed on SMEs.

Alternatively, weights for test specifications can be derived using an actuarial or statistical model (Cornelius & Lyness, 1980; Kane, et al., 1989; Lunz, et al., 1989). Statistical models allocate weights to domains based on the results of a task analysis and other relevant data. The use of a statistical model implies that there are advantages that might not be realized through the use of holistic SME judgments. First, there is ample evidence drawn from a variety of contexts indicating that the expectations of experts often do not hold up to empirical scrutiny. The process of translating task analysis data into test specifications requires SMEs to evaluate and integrate vast amounts of information. This is exactly the type of circumstance that may produce inaccurate judgments (Dawes, Faust, & Meehl, 1989). Furthermore, problems of SME unreliability have been documented specifically within the context of job analysis and the development of test specifications (Hughes and Prien, 1989). Second, test specifications have the potential of reflecting the biases of SMEs, which in turn, may limit the generalizability of test specifications (Nelson, 1994). Conducting an empirical task analysis is just one step in the direction of replicability. By complementing an empirical task analysis with systematic procedures for collecting and combining relevant judgments, the likelihood of replication can be increased. Third, the use of a statistical model might assist in accomplishing certain objectives related to sound measurement practice. More, specifically, professional guidelines indicate that it is important to establish a clear link between job responsibilities and the KSAs to be tested on an examination (AERA, et al., 1985; D'Costa, 1986; Levine, Thomas, & Sistrunk, 1988; Primoff & Fine, 1988; Thompson & Thompson, 1982). The use of a statistical model is one way to unambiguously describe such a link. The goal of a statistical model, then, is not to supplant human judgment, but rather to elicit, quantify, and integrate those judgments in a systematic manner.

Statistical models vary according to the factors included in the model (i.e., the attributes of the task that are of interest), the emphasis allocated to each factor, and the mathematical form of the model. For example, a model might state that a task's overall importance is determined by obtaining the product of the frequency, time spent and criticality of that task. Some researchers have offered convincing arguments in support of multiplicative models (Kane, et al. 1989), while others have provided convincing data supporting the use of simple linear models (Sanchez & Levine, 1989). The Rasch model has also been proposed for use in establishing weights for test specifications (Lunz, et al., 1989). Meanwhile, some studies indicate that different models yield similar results (Sanchez & Fraser, 1992; Schmeiser, 1987; 1988). In short, research on the form

and utility of statistical models is mixed and inconclusive. There is one conclusive point, and that point concerns the importance of differentiating between nominal weights and effective weights when formulating such models (Schmeiser & Estes, 1986). One study found, for example, that frequency ratings were much more variable than criticality ratings, which meant that frequency would exert more influence in the model than criticality, unless the criticality data were first transformed (Kane, et al., 1989). The effective contribution of a variable depends on the variances and covariances of all the variables included in the model. If variances and covariances are not taken into consideration, then the model will not be doing what it was intended to do.

The use of a statistical model assumes that such a model will provide meaningful results. It could very well be that the process of establishing weights for test specifications is too complex to capture with a relatively simple model, or that other factors preclude the utility of statistical models. Although prior research is limited, there have been several studies at least related to the matter of eliciting and combining judgments within the context of job analysis. Cornelius and Lyness (1980) found that a statistical algorithm which combined detailed information about tasks and KSAs provided more consistent results than holistic judgments. That study also evaluated a method that required SMEs to subjectively combine the same type of detailed ratings used in the statistical model. This latter method was substantially inferior to the holistic judgments and statistical model.

Another study compared holistic judgments to statistical weights for four certification examinations in radiologic technology (Raymond, 1994). The statistical model included ratings of task frequency, task criticality, and information concerning the task – KSA linkages. First, task frequency and task criticality ratings were combined into a single index of importance, using a simple additive model that gave criticality twice the effective weight of frequency. Next, the importance ratings were integrated with the results of a linkage matrix completed by a panel of SMEs. The statistical model was specified so that the magnitude of the weight assigned to a KSA domain on the test specifications was a function of the number of tasks linked to that domain and the importance of those tasks. Holistic weights for each KSA domain were also obtained. They were provided by the same panel of SMEs that completed the linkage exercise. Although the results of this study were mixed, there was a consistent pattern of agreement and disagreement across the test specifications for the four certification programs. The two methods exhibited moderate to high levels of agreement for part of the test specifications that addressed specific imaging procedures, but showed very little agreement for more general sections of the test specifications. One possible problem was the use of an overly simplistic statistical model, which required only dichotomous linkage judgments (present or absent). A polytomous scale may have been more suitable. The quality of the task statements and format of the test specifications also appeared to be important. In general, weights based on holistic judgments were similar to weights obtained from the statistical model when the tasks were unambiguous, nonoverlapping, and linked with procedurally-oriented categories. Do these results mean that test specifications should be organized in terms of procedures?

Organizational Frameworks Revisited

Test specifications can be erected on two types of foundations, often referred to as test content and examinee process. Test content usually refers to topics, subjects, or knowledge domains. *Descriptors of test content are often nouns.* The process component usually refers to the type of performance required of the

examinee. Processes may include cognitive skills (e.g. analyze) or job functions (e.g., diagnose). *Descriptors of processes are often verbs.* If a verb is not used, then the nominal form of a verb is often used (e.g., written comprehension instead of comprehend written text). Content and process information can be combined to create a variety of frameworks, as noted below.

- An outline of test content. Such and outline would typically list the topics included on a test. The outline might be very specific and lengthy or very general and short.
- An outline of examinee processes. Such an outline might include tasks, procedures, or cognitive processes for which examinees are expected to demonstrate knowledge or proficiency. The outline might be very specific and lengthy or very general and short.
- A content-by-process matrix. The matrix could be completely or incompletely crossed.
- Content nested within processes
- Processes nested within content
- An outline consisting of a mixture of content categories and process categories. Either content or process could be comprised of one or more dimensions.

For many years, it seemed as if test specifications were organized in much the same way as textbooks or curricula; that is, test specifications truly resembled a content outline. The obvious advantage of content-oriented test specifications is that many content domains were well mapped, logically structured, and widely used, thereby allowing test developers, SMEs, and examinees to speak a common language. Furthermore, the organization of test content often resembled the way examinees learned that content. Section scores were probably regarded as quite meaningful by examinees. Then, for a variety of reasons, many test specifications began to include a process component, or were completely organized around practice-based performance domains. Although it is difficult to say exactly what spawned this interest, several events occurred during that era. First, there was an orchestrated movement underway to upgrade credentialing in the health professions. This upgrade included balancing the importance of curriculum with the results of empirical job analyses (National Commission of Health Certifying Agencies, 1981). In addition, process models such as Bloom's Taxonomy were becoming more popular, and there was a growing interest in building tests that evaluated application of knowledge rather than recognition of factual information. The use of patient management problems was also increasing. Problem-based learning, case studies, and other instructional methods that integrated subject-matter knowledge with problem-solving strategies were also being promoted. The nursing process is one example of a process model that formed the basis of many content outlines. Other examples can be found in medicine (e.g., Levine, et al., 1968; Miller, 1968).

Process-based test specifications usually have a direct and obvious relationship to practice or to other real-world tasks. This is a desirable property. The obvious limitation is that no occupation is content free. All test questions, even those that require complex cognitive processes, also deal with some type of subject matter (Perkins & Salomon, 1989). It is responsibility of credentialing agencies to assure that candidates have mastered that subject matter. Furthermore, if test specifications do not specify subject matter, then it is difficult to write appropriate test questions and develop comparable test forms. Another limitation of process-based outlines is that variation in test performance is more likely to be influenced by test content than by process. People usually miss test questions because they don't know the content (e.g., renal system, disc brakes,

generalizability theory) not because can't diagnose, analyze, or problem solve. This fact has been repeatedly documented and is referred to as content specificity (Elstein, Shulman, & Sprafka, 1978). Section scores derived from process models have little psychological meaning or empirical utility. The significance of this limitation becomes apparent when candidates who fail an exam ask, "I scored lower in planning, what should I study?" or "I scored lower in diagnosis, can you recommend a book for me?"

Many credentialing exams are now built around a content-by-process matrix, thereby overcoming the limitations associated with the use of just one or the other. A discussion of process-by-content matrices and is offered Millman & Greene (1989). Lunz et al. (1989) provides a nice example of a content-by-process matrix for laboratory managers. The content dimension consists of primarily of various management disciplines such as financial management, operations management, and personnel management. The process dimension consists of management tasks, and is comprised of categories such as applies, calculates, and develops. Each test item can be classified into one category for each of the two dimensions. Multidimensional matrices are also common, and have been advocated as very effective means for describing practice in various occupations (D'Costa, 1986, LaDuca et al., 1984; McGaghie, 1980).

The obvious advantage of content-by-process matrices is that they not only indicate the content, but also specify what the examinee is expected to do with that content. They also encouraging item writers and test developers to focus on job-related processes and skills. Each cell of a multidimensional practice model prescribes specific clinical situations, which can be useful for providing item writers with the necessary raw materials for generating questions (D'Costa, 1986, LaDuca et al., 1984. McGaghie, 1980). Another advantage is that examinees can be given feedback on both content and process dimensions. Content-by-process matrices do not solve all testing needs, and test developers must still overcome a few practical obstacles. One consideration is that a multidimensional model will produce many cells — sometimes in the thousands (McGaghie, 1980). It may not be possible to have test items for every cell on each form of an exam. This can be addressed by specifying numbers of questions on the margins of the matrix, rather than for each cell. Alternatively, the test specifications might indicate the probability that a test item addressing a certain cell will be included on the test. A second consideration is that certain test content on one dimension may cut across multiple levels of another dimension. For example, a question on neuroanatomy might apply to several process categories. Again, working on the margins provides one way to manage such concerns. Another solution would be to create a completely independent section of the test specifications that deal with fundamental knowledge and basic principles. For example, Section I of an engineering exam might consist of basic sciences (math and physics), while Section II would consist of a multidimensional problem-based matrix. Such test specifications are relatively common in credentialing. A third consideration has to do with usability. Multidimensional frameworks can be complicated and cumbersome; attention must be given to assuring that they can be readily comprehended by relevant audiences.

Organizing test specifications is a matter of matching an organizational framework with the type of exam being administered. It should be evident that a content-by-process matrix can be made to work for most types of exams. If a content-by-process matrix is deemed unnecessary or undesirable, then a choice must be made between content or process. Most credentialing exams consist of multiple-choice questions (Shimberg, 1981), which emphasize what people know. For such exams, test specifications organized around content are likely to

serve their intended purposes. For performance-based credentialing exams that emphasize what people can do (e.g., orals or work samples), a set of test specifications organized around process will likely be most effective.

Evaluating the Effectiveness of Job Analysis

One purpose of this symposium is to examine the extent to which licensure and certification testing is serving the needs of the ***public*** who rely on the services of credentialed individuals, the ***examinees*** seeking the credential, and the ***credentialing agencies*** that administer the licensure or certification program. Given that job analysis plays a crucial role in test validity, it is important to ask whether job analysis as we know it really works? The short answer is, well, probably. Although the preceding discussion dwelled on problems and challenges, current methods of job analysis are likely more than adequate. Can we do better? The answer is yes, and one of the primary purposes of this paper has been to suggest avenues for improvement. Issues related to this are discussed below.

Protecting the Public

Credentialing exams protect the public by identifying individuals who are qualified to practice, or individuals who are entitled to use a particular job title or credential. Job analysis and test specifications serve the public interest in at least two ways[4]. First, they identify tasks and KSAs that are required in order to protect the public. A job analysis that is deficient will fail to identify the important KSAs, thereby increasing the risk of exposing the public to a practitioner that is not fully qualified. In practice, most job analyses are so comprehensive that the public is generally well served. Even in the absence of a job analysis, a group of SMEs could probably meet and easily specify the KSAs required of a credentialed individual.

The second way that job analysis serves the public is by limiting the test specifications to only those KSAs required for safe and effective practice. If the job analysis and test specifications are contaminated by tasks and KSAs that are not truly relevant, then the public may be adversely affected. One potential problem is that the presence of irrelevant KSAs on an exam might displace other relevant KSAs. Another potential problem is that the public's choice of services can be restricted if significant numbers of candidates fail an exam due to the presence of irrelevant content. The restriction could take the form of increased prices or limited access to services. It is not difficult to imagine the serious negative social consequences that could result from excluding qualified practitioners from the workforce.

Job analysis has earned a mixed report card in terms of meeting this second obligation. On one hand, most methods of job analysis will prevent irrelevant tasks from finding their way onto test specifications. On the other hand, current methods for linking task performance to KSA requirements are often vague and ambiguous, resulting in tenuous links (Nelson, 1994). The process of establishing task–KSA linkages and translating job analysis results into test specifications is in need of refinement and standardization (e.g., Levine, et al., 1980; Shimberg, 1990). Another aspect of job analysis in need of additional work is that of specifying the

[4] This discussion assumes that a licensing exam is truly needed to protect the public. Parts of this discussion may not apply to voluntary certification exams.

level of proficiency or depth of knowledge associated with each KSA. This is necessary to ensure that the level of performance required for each KSAs is not higher than the level required in practice. The types of analyses required to guarantee this are currently absent from most job analyses.

Protecting Examinees

A credentialing examination has numerous responsibilities to examinees regarding the way exams are developed, administered, and scored. The consequences of a poorly developed or unfair exam are obvious: a candidate who has trained for an occupation or profession, sometimes for years, might be denied the opportunity to practice.

Job analysis has an important, but limited, role in protecting examinees from unfair testing practices. First and foremost, test specifications must be developed in a manner that protects examinees from test content that is not relevant to practice. Test specifications should also be of sufficient detail so that examinees can identify the content and prepare for examinations. Well-organized test specifications can also facilitate remediation efforts, by providing the basis for useful section scores. Just as thorough job analyses have served the public well, I think that examinees have, for the most part, been assured of relevant credentialing exams through the efforts of job analysis. However, as argued earlier, procedures for establishing task–KSA linkages and for specifying levels of proficiency are in need of refinement. It could also be argued that methods for providing feedback to examinees could be improved.

Protecting Credentialing Agencies

There are two levels of protection with which a credentialing agency is concerned. At one level, they are responsible to the public, to examinees, and perhaps to other constituencies (e.g., the profession). Job analysis protects credentialing agencies to the extent that the public, examinees, and other relevant constituencies are protected. Credentialing agencies should have a particularly keen interest in job analysis as the use of computers to assemble and administer tests becomes increasingly common. As noted earlier, if computers are to be effective at testing, then it is essential that test specifications be well-organized and precise (e.g., Swanson, Stocking, & Pearlman, 1993). In particular, attempting to guarantee the content representation of a short computer-adaptive test will place great demands on test specifications and item classification.

At another level, credentialing agencies are concerned with avoiding litigation (Smith & Hambleton, 1990). Are current methods for conducting job analysis serving this interest? One might contrive two somewhat inconsistent lines of thought on this matter. The optimistic line of thought goes something like this:

> *Job analysis appears to be serving credentialing agencies admirably. The fact that a job analysis has even been conducted is probably sufficient for meeting legal obligations. As noted in the Uniform Guidelines), "any method of job analysis may be used if it provides the information required for the specific validation strategy used" (EEOC et al., 1978).*

A more challenging view might go something like this:

> *There is a growing consensus that a job analysis should include evidence of the actual tasks performed (D'Costa, 1986; Levine, Thomas & Sistrunk, 1988; Primoff & Fine, 1988; Thompson & Thompson, 1982). Furthermore, the courts have indicated that it is important to demonstrate a link between actual tasks and the content of the test (Kuehn, Stallings, & Holland, 1990).*

One interpretation of the first point would suggest that the status quo is adequate, and that the interest of credentialing agencies is indeed protected by current methods of job analysis . The second point suggests that "demonstrating a link" is important. The second point implies that there is room for improvement, since linkage efforts are not always performed, and since the methods for establishing linkages are not well understood. The *Standards* (AERA, et al, 1985) are not particularly helpful in interpreting these issues, or in providing concrete guidelines for conducting job analysis (Shimberg, 1990).

The Science and Politics of Validity

Validity refers to the degree to which empirical evidence and theoretical rationales support the inferences and actions based on test scores (Messick, 1989). The opening paragraph of this paper stated the obvious: that job analysis is the foundation upon which valid credentialing exams are built. The link between job analysis and test validity is firmly imprinted in everyone's mind. However, even though a job analysis is necessary for the development of valid credentialing examinations, it is far from a sufficient condition. One obvious reason is that there are many sources of variation in test scores that have very little to do with whether or not a job analysis is conducted (see Stanley, 1972 for one list of sources). A second reason is that job analysis represents only one limited component of the testing enterprise. Activities that occur subsequent to the job analysis include the development of test specifications, item writing, standard setting, and equating, to name a few. Oversights in any of these critical activities can interfere with test validity. This paper has been limited to a discussion of the manner in which job analysis and the development of test specifications can impact validity. This paper has not discussed the numerous other factors that must be considered when interpreting test scores. There are many. For example, job analysis does not:

- assure that passing scores are set at appropriate levels (i.e., levels that maximize the accuracy of pass/fail decisions).
- tell us how to combine data from different assessment methods (e.g., orals and written exams) in order to maximize the accuracy of pass/fail decisions.
- prevent or detect item or test bias.
- assure that test questions measure primarily relevant KSAs instead of irrelevant constructs (e.g., reading skill for written tests or interpersonal skills for oral exams).

The list goes on and on. Although job analysis supports test development efforts, conforms to professional standards, and fulfills legal requirements, it tells us very little about the nature of test scores. In short, job analysis satisfies the political side of validity. But there is so much more. At the same time that the measurement community is investigating methods to improve job analysis and the development of test specifications, we need to be expanding our strategies for validating scores resulting from the administration of credentialing examinations. It is time to also foster the science of validating credentialing examinations.

References

ACT (1991). *Results of the job analysis and content validity study for Certified Professional Secretaries.* Kansas City, MO: Professional Secretaries International.

ACT (1992). *The CCNA practice analysis and test specifications project.* Park Ridge, IL: Council on Certification of Nurse Anesthetists.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington DC: American Psychological Association.

Ash, R.K. (1988) Job analysis in the world of work. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government, Vols. I and II* (pp. 3-13). New York, NY: Wiley & sons.

Butler, S.K. & Harvey, R.J. (1988). A comparison of holistic versus decomposed rating of Position Analysis Questionnaire work dimensions. *Personnel Psychology, 41,* 761-771.

Cascio, W.F. (1982). *Applied psychology in personnel management,* Reston, VA: Reston Publishing.

Cizek, G.J., Webb, L.C., & Kalohn, J.C. (1995). The use of cognitive taxonomies in licensure and certification test development: reasonable or customary? *Evaluation and the Health Professions, 18,* 77-91.

Cornelius, E.T. (1988). Practical findings from job analysis research. In S. Gael (Ed.) *The job analysis handbook for business, industry, and government, Vols. I and II* (pp. 48-64). New York, NY: Wiley & sons.

Cornelius, E.T. & Lyness, K.S. (1980). A comparison of holistic and decomposed judgment strategies in job analysis by job incumbents. *Journal of Applied Psychology, 65,* 155-163.

Cranny, C.J. & Doherty, M.E. (1988). Importance ratings in job analysis: Note on the misinterpretation of factor analysis. *Journal of Applied Psychology, 73,* 320-322.

Dawes, R.M., Faust, D. & Meehl, P. (1989). Clinical versus actuarial judgment, *Science, 243,* 1668-1674.

D'Costa, A. (1986). The validity of credentialing examinations. *Evaluation and the Health Professions, 9,* 137-169.

Dunnette, M. D. (1976). Aptitudes, abilities, & skills. In M.D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 473-520). Chicago, IL: Rand-McNally.

Edwards, J.C., Currie, M.L., Wade, T.P., & Kaminski, D.L. (1993). Surgery resident selection and evaluation: a critical incident study. *Evaluation and the Health Professions, 16,* 73-86.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1978). *Medical problem solving: an analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of uniform guidelines of employee selection procedures. *Federal Register, 43*(166), 38290-38315.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Fleishman, E.A. (1975). Toward a taxonomy of human performance. *American Psychologist, 30,* 1127-1149.

Fleishman, E.A. (1982). Systems for describing human tasks. *American Psychologist, 37,* 821-834

Fleishman, E.A. & Quaintance, M.K. (1984). *Taxonomies of human performance: the description of human tasks.* New York, NY: Academic Press.

Fidler, J. (1988). Medical assistants: an analysis of tasks performed in practical settings. *Evaluation and the Health Professions, 11,* 358-378.

Fine, S. (1986). Job analysis. In R.A. Berk (Ed.), *Performance assessment: Methods and Applications.* Baltimore, MD: Johns Hopkins University Press.

Friedman, L. (1990). Degree of redundancy between time, importance and frequency task ratings. *Journal of Applied Psychology, 75,* 748-752.

Gael, S. (1983). *Job analysis: a guide to assessing work activities.* San Francisco, CA: Jossey-Bass.

Gael, S. (1988). *The job analysis handbook for business, industry, and government, Vols. I and II.* New York, NY: Wiley & sons.

Hubbard, J.P., Levit, E.J., Schumacher, C.F., & Schnabel, T.G. (1965). An objective evaluation of clinical competence. *New England Journal of Medicine, 272,* 1321-1328.

Hughes, G.L. & Prien, E.P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. *Personnel Psychology, 42,* 283-292.

Iowa Intervention Project (1993). The NIC taxonomy structure. *Image, 25*(3), 187-192.

Kane, M.T. (1982). The validity of licensure examinations. *American Psychologist, 37,* 911-918.

Kane, M.T., Kingsbury, C. Colton, D., & Estes, C. (1986). A study of nursing practice and role delineation and job analysis of entry-level performance of registered nurses. Chicago, IL: National Council of State Boards of Nursing.

Kane, M.T., Kingsbury, C. Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing plans for licensure and certification examinations. *Journal of Educational Measurement, 26,* 17-27.

Kuehn, P.A., Stallings, W.M., & Holland, C.L. (1990). Court-defined job analysis requirements for validation of teacher certification tests. *Educational Measurement: Issues & Practice, 9*(4), 21-24.

Kunen, S., Cohen, R., & Solman, R. (1981). A levels of processing analysis of Bloom's taxonomy. *Journal of Educational Psychology, 73,* 202-211.

LaDuca, A. (1980). The structure of competence in the health professions. *Evaluation and the Health Professions, 3,* 253-288.

Laduca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation and the Health Professions, 17,* 178-197.

LaDuca, A. Engle, J.D., & Risley, M.E. (1978). Progress toward development of a general model for competence definition in the health professions. *Journal of Allied Health, 7,* 149-155.

LaDuca, A., Taylor, D.D., & Hill, I.K. (1984). The design of a new physician licensure examination. *Evaluation and the Health Professions, 7,* 115-140.

Landy, F.J. (1988). Selection procedure development and usage. In S. Gael (Ed.) *The job analysis handbook for business, industry, and government, Vols. I and II* (pp. 271-287). New York, NY: Wiley & sons.

Landy, F.J. & Vasey, J. (1991). Job analysis: the composition of SME samples. *Personnel Psychology, 44,* 27-50.

Levine, E.L., Ash, R.A., & Bennett, N. (1980). Exploratory comparative study of four job analysis methods. *Journal of Applied Psychology, 65,* 524-535.

Levine, E.L., Thomas, J.N., & Sistrunk (1988). Selecting a job analysis approach. In S. Gael (Ed.) *The job analysis handbook for business, industry, and government, Vols. I and II* (pp. 339-352) New York, NY: Wiley & sons.

Levine, H.G., McGuire, C., Miller, G.E., & Larson, C. (1968). *The orthopaedic training study: Final report.* Chicago, IL: University of Illinois, Center for the Study of Medical Education.

Lunz, M.E., Stahl, J.A., & James, K. (1989). Content validity revisited: transforming job analysis data into test specifications. *Evaluation and the Health Professions, 12,* 192-206.

McCloskey, J.C. & Bulachek, G.M., Eds. (1992). *Nursing Interventions Classification,* St. Louis, MO: Mosby-Year Book.

McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire. *Journal of Applied Psychology, 56,* 347-368.

McGaghie, W.C. (1980). The evaluation of competence: Validity issues in the health professions. *Evaluation and the Health Professions, 3,* 289-320.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational Measurement* (pp. 13-103. New York, NY: American Council on Education and Macmillan.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.) *Education measurement (3rd ed.).* New York, NY: Macmillan.

Miller, G.E. (1968). The orthopaedic training study. *Journal of the American Medical Association, 206,* 601-606.

National Commission of Health Certifying Agencies (1981). *Task force report on education and certification.* Washington, DC: Author

National Council of Examiners for Engineering and Surveying (1989). *Analysis of professional activities and requirements of the engineering profession.* Clemson, SC: Author.

Nelson,. D. (1994). Job analysis for licensure and certification exams: science or politics. *Educational Measurement: Issues & Practice, 13* (4), 29-35.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin, 87,* 1-28.

Perkins, D.N. & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*(1), 16-25.

Peterson, N. G. & Bownas, D.A. (1984). Skill, task structure, and performance acquisition. In M.D. Dunnette and E.A. Fleishman (Eds.), *Human performance and productivity, Vol 1: Human capability assessment* (pp. 49-105.

Primoff, E.J. & Fine, S.A. (1988) A history of job analysis. In S. Gael (Ed.) *The job analysis handbook for business, industry, and government, Vols. I and II* (pp. 14-29). New York, NY: Wiley & sons.

Raymond, M.R. (1989). Applications of multidimensional scaling to research in the health professions. *Evaluation and the Health Profession, 12,* 379-408.

Raymond, M.R. (1994, April). *Equivalence of weights for test specifications obtained using empirical and judgmental procedures.* Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Raymond, M.R., Buckwalter, J.A., & Steyers, C.M. (1989, March). *Investigating the structure of orthopedic proactive using multidimensional scaling analysis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rodgers, J.L. (1988). Structural models of the American Psychological Association in 1986: a taxonomy for reorganization. *American Psychologist, 43,* 372-382.

Rosenfeld, M., Shimberg, B., & Thornton, R.F. (1983). *Job Analysis of Licensed Psychologists in the United States and Canada.* Princeton, NJ: Educational Testing Service.

Sanchez, J.I. & Fraser, S.L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology, 77,* 545-553.

Sanchez, J.I. & Levine, E.L. (1989). Determining important tasks within jobs: a policy-capturing approach. *Journal of Applied Psychology, 74,* 336-342.

Schafer, L., Raymond, M.R., & White A.S. (1992). A comparison of two methods for structuring performance domains. *Applied Measurement in Education, 5,* 321-335.

Schmeiser, C.B. (1988, April). *Effects of statistical and judgmental weightings of task analysis data on test specifications.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Schmeiser, C.B. (1987, April). *Effect of translating task analysis into test specifications.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC.

Schmeiser, C.B. & Estes, C.A. (1986, April). *Translating task analysis results into test specifications: clarity or confusion?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Seddon, G.M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research, 48,* 302-323.

Sireci, S.G. & Geisinger, K.F. Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16,* 17-31.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36,* 1138-1146.

Shimberg, B. (1990). Social considerations in the validation of licensing and certification examinations. *Educational Measurement: Issues and Practice, 9*(4), 11-14.

Smith, I.L. & Hambleton, R.K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues & Practice, 9*(4), 7-10.

Sokal, R. (1974). Classification: Purpose, principles, progress. *Science, 185,* 1115-1123.

Stanley, J.C. (1971) Reliability. In R.L. Thorndike (ed.), *Educational measurement, 2nd ed.*. Washington, DC: American Council on Education.

Stocking, M.L., Swanson, L., & Pearlman, M. (1993). Application of item selection method to real data. *Applied Psychological Measurement, 17,* 167-176.

Tannenbaum, R.J. & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: a study in the context of licensure testing. *Journal of Applied Psychology, 78,* 975-980.

Tarrico, V.S., Smith, W.L., Altmaier, E.M., Franken, E.A., & Van Velzen, D.V. (1984). Critical incident interviewing in evaluation of resident performance. *Radiology, 152,* 327-329.

Thompson, D.E. & Thompson, T.A. (1982). Court standards for job analysis in test validation. *Personnel Psychology, 35,*865-874.

Wilson-Pessano, S. (1988, April). *Defining professional competence: the critical incident technique 40 years later.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Zaglaniczny, K.L. (1993). Council on Certification Professional Practice Analysis. *Journal of the American Association of Nurse Anesthetists, 61,* 241-255.